On the Reliability of Software Piracy Statistics

Abstract

Based on a review of the methodology and empirical analysis of software piracy statistics compiled by the Business Software Alliance (BSA), I conclude that a change in the BSA consultant and methodology around 2002-03 had systematic effects on published piracy rates. The trend rate of decrease of piracy rates fell from 2.0% points per year to 1.1% points per year. Among countries for which the BSA's piracy statistics depended on a projection of software usage, the trend decrease in the piracy rate was under-estimated and the sensitivity of piracy to changes in income was over-estimated. Any government pronouncement or action, and academic study using the BSA software piracy statistics should take account of changes in the BSA consultant and methodology.

1. Introduction

Special 301 of the U.S. Trade Act of 1974 requires the U.S. Trade Representative to report annually on countries that do not provide adequate and effective protection of intellectual property (IP) rights, or deny fair and equitable market access to U.S. exporters of IP-protected items. Countries whose laws, policies, or practices are deemed to adversely affect U.S. producers or products may be subject to investigation, trade sanctions, or other penalties. Piracy statistics reported by the International Intellectual Property Alliance (IIPA) are the key evidence in Special 301 reviews.

The public-policy implications of piracy motivated both the World Intellectual Property Organization and the Organization for Economic Cooperation and Development to commission studies to measure piracy [12]. As well, numerous academic researchers have investigated the causes of business software piracy [1][4][5][6][7][8][9][11][13][14][15].

Here, I review the methodology, coverage, and implementation of the BSA software piracy statistics and previous studies of the causes of software piracy. Based on various econometric tests, I conclude that a change in the BSA consultant and methodology around 2002-03 had systematic effects on published piracy rates. First, the trend rate of decrease of piracy rates fell from 2.0% points per year to 1.1% points per year. Second, among countries for which the BSA's piracy statistics depended on a projection of software usage, the trend decrease in the piracy rate was under-estimated and the sensitivity of piracy to changes in income was over-estimated.

The central implication of my analysis is that BSA statistics should be used with great caution. Any government pronouncement or action, and academic study using the BSA software piracy statistics should take account of changes in the BSA consultant and methodology. Software piracy statistics for countries for which BSA undertook projections of software usage should be treated with special care.

2. Methodology

For 2002 and earlier years, the business software piracy statistics were produced by the International Planning and Research Corporation (IPRC) [3]. IPRC focused on three groups of business PC software – general productivity applications, professional applications, and utilities.

The IPRC estimated piracy using an indirect methodology. For each country, the quantity of pirated software was estimated as being the difference between the quantity installed and the quantity legitimately acquired. In turn, the quantity installed was estimated as the number of computers in use multiplied by corresponding norms for the "software load" in four customer segments -- new and existing residential computers, and new and existing business computers.

Software load is the quantity of software installed per computer. The norms for software load for the four segments were based on U.S. market research ([3] page 11). However, the IPRC did not explain whether, and if so, how it adjusted the U.S. norms to compute the software load in other countries.

The IPRC directly estimated the numbers of computers in use "for the major countries … from proprietary and confidential data supplied by BSA member companies", while "[t]he "rest of region" data was used to develop piracy estimates outside of the major markets" ([3] pp. 11-12). The IPRC did not specify the "major" markets or method by which it developed the "rest of region" data.

In 2003, BSA engaged a new consultant, International Data Corporation (IDC), and the methodology underlying the business software piracy statistics was refined. IDC expanded the scope of measurement to cover all PC software, including operating systems, systems software such as databases and security packages, and general and specific applications software.

The IDC applied the same basic methodology, estimating the quantity of pirated software indirectly as the difference between the quantity installed and the quantity legitimately acquired. In turn, the quantity installed was estimated as the number of computers in use multiplied by norms for the software load.

IDC calculated the software loads per new computer and per existing computer from surveys of consumers and business users in 15 countries. IDC explained that the "results of these surveys were used to populate IDC's input models for the other countries. However, IDC did not explain whether, and if so, how it adjusted the norms from the 15 countries to compute the software load in the other countries.

As for the numbers of computers, IDC collected information on PC shipments for "more than 75 countries", while for the "additional 25-plus countries and markets, the data were either collected in-country or modeled regionally based on IDC's rest-of-region estimates" ([2], page 10). The IDC did not explain the modeling.

3. Average Piracy Rates

To study the impact of the change in consultant and methodology, I compiled national piracy rates over the period 1997-2007 from BSA publications. The period of study included seven "pre-change" years (1997-2002) and five "post-change" years (2003-07). The panel began with 81 countries in 1997 and ended with 103 countries in 2007.

Figure 1 illustrates the evolution of the average piracy rate as published by the BSA, in red. Evidently, the average piracy rate declined by about 2% points each year until 2002. Between 2002 and 2003, the average piracy rate jumped by over 2% points, and thereafter, continued a downward trend, but at a much lower rate of decline than before 2003.

The break in the trend decrease of piracy coincided exactly with the change in consultant and methodology. Table 1 reports various econometric tests to confirm the impact on national piracy rates. Specifications (i) and (ii) were simple ordinary least-squares (OLS) regressions of the piracy rate on year indicators and pre- and post-change year trends respectively. The results from both specifications suggest that piracy rates followed a significant downward trend.

Referring to specification (i), the year indicators show a clear decreasing trend in the average piracy rate from 1999-2002. The results from specification (ii) are even clearer: prior to the change in consultant and methodology, the average piracy rate fell by 2.0% points annually, while, after the change, the average rate fell at the slower rate of 0.7% points annually. The difference between the pre-change and post-change trends was statistically significant ($F(1, 989) = 15.30$, $p < 0.0001$).

The OLS regression pooled all countries, regardless of economic, institutional, or cultural differences, into a single estimate. Obviously, it would be more appropriate to account for any systematic national differences. Specifications (iii) and (iv) replicated the analysis, using country fixed effects. The fixed effects would account for any systematic national differences which did not vary over time. The results were even stronger than with the OLS estimates.

Referring to specification (iii), the year indicators show a clear decreasing trend from 1998-2002, and then an upward shift by about 1.7% points between 2002-03. Referring to specification (iv), the trend rate of decrease of piracy rates was significantly higher before the change in consultant and methodology (–2.0% points per year) than after the change (–1.1% points per year). The difference between the pre-change and post-change trends was statistically significant ($F(1, 102) = 60.18$, $p < 0.0001$).

One possible reason why the downward trend of piracy rates decelerated around 2002-03 was the expansion of BSA coverage to include countries with higher piracy rates. Specifically, in 1997, the BSA piracy statistics expanded coverage from 81 countries in 1997, to 103 countries in 2007. The expansion included countries with relatively high piracy.

To avoid any bias due to the expanded coverage, specifications (v) and (vi) limited the fixed effects estimates to those countries covered throughout the period, 1997-2007. The results from the balanced sample were similar. Referring to specification (vi), the pre-change trend rate of decrease of piracy rates (–2.1% points per year) was significantly higher than the post-trend change (–1.1% points per year) ($F(1, 80) = 58.55$, $p < 0.0001$).

To give context to this change in trend, in Figure 1, I add in blue, the projection of the average piracy rate, based on the year trend, as reported in Table 1, column (vi). In 2007, the average piracy rate, as published by the BSA, was just over 55%. If the average piracy rate had continued its pre-2002 downward trend, then, by 2007, it would have been about 45%, which is a substantial disparity from the published rate.

4. Projection by Income

Formally, the methodology applied by BSA consultants, IPRC and IDC, was to estimate the quantity pirated in country *i* during year *t* as the difference between usage of software, U_{it}, and the quantity of software legitimately acquired, S_{it},

$$P_{it} = U_{it} - S_{it} . \quad (1)$$

In essence, the usage was computed as

$$U_{it} = \lambda_i N_{it} , \quad (2)$$

where λ_{it} was the norm for software load and N_{it} was the number of computers in country *i* in year *t*.

The piracy rate in country *i* for year *t* was then calculated as the ratio of the pirated quantity to usage,

$$r_{it} = \frac{P_{it}}{U_{it}} = \frac{U_{it} - S_{it}}{U_{it}} = 1 - \frac{S_{it}}{U_{it}} . \quad (3)$$

As noted above, both the IPRC and IDC measured the software load and number of computers in particular sample countries, and estimated the software load and number of computers in other countries. The interesting and important question is how they generated these estimates.

Multiple academic studies have pointed to income as being the single most important influence on the rate of software piracy [5][6][7][9][11][14]. Referring to equation (3), these observations are consistent with income being an important determinant of both legitimate sales and usage. This theoretical analysis leads to the following:

Hypothesis. Software load or number of computers was estimated on the basis of national income per capita.

To test this hypothesis, I used instrumental variable methods. First, I used equation (3) to obtain the rate of legitimate consumption as

$$c_{it} = 1 - r_{it} = \frac{S_{it}}{U_{it}} . \quad (4)$$

In logarithms, this becomes

$$\ln c_{it} = \ln S_{it} - \ln U_{it} . \quad (5)$$

Referring to (2), under the null hypothesis, the BSA estimated software load or number of computers by income. Accordingly, suppose that the reduced form for software usage is an increasing function of income,

$$U_{it} = Y_{it}^{\alpha} , \quad (6)$$

with $\alpha > 0$. Then, substituting from (6) in (5), the legitimate consumption rate (in logarithm) as estimated by BSA would be

$$\ln \tilde{c}_{it} = \ln S_{it} - \alpha \ln Y_{it} . \quad (7)$$

Consider a regression of $\ln \tilde{c}_{it}$ on $\ln Y_{it}$,

$$\ln \tilde{c}_{it} = \ln S_{it} - \alpha \ln Y_{it} = \beta_1 \ln Y_{it} + \beta_2 X_{it} + \varepsilon_{it} , \quad (8)$$

where β_1 and β_2 are coefficients, and X_{it} represents other covariates which might affect legitimate consumption. Under the null hypothesis, in (8), the error term, ε_{it}, would be correlated with the explanatory variable, $\ln Y_{it}$. Equivalently, income would not be exogenous. The essential reason is the projection of software usage on the basis of income.

A Hausman test ([16], pp. 532-533) of the endogeneity of income in (8) would proceed as follows. Perform first stage regressions for every income variable – regressing every income variable on instruments for the income variables, Z_{it}, and exogenous variables, X_{it}, that affect piracy.

$$\ln Y_{it} = \gamma_1 X_{it} + \gamma_2 Z_{it} + \eta_{it} , \quad (9)$$

where γ_1 and γ_2 are coefficients. Extract the residuals from the first-stage regression,

$$u_{it} = \ln Y_{it} - \gamma_1 X_{it} - \gamma_2 Z_{it} . \quad (10)$$

In the second stage, regress the rate of legitimate consumption on the income variables, exogenous variables that affect piracy, and the residuals from the first-stage regressions,

$$\ln \tilde{c}_{it} = \theta_1 \ln Y_{it} + \theta_2 X_{it} + \theta_3 u_{it} + \upsilon_{it} , \quad (11)$$

where θ_1, θ_2, and θ_3 are coefficients. If the first-stage residuals are significant in the second-stage regression, then the corresponding first-stage regressor is not exogenous. The implication would be that, indeed, legitimate consumption and hence piracy rates were based on income.

Tables 2 and 3 present descriptive statistics of the data and pairwise correlations respectively. The key issue in any instrumental variables analysis is the choice of instrument. I identified per capita residential electricity consumption as an instrument. Residential electricity consumption would depend on income, but, intuitively, there would not seem to be any direct relation between electricity consumption and legitimate consumption of software, except to the trivial extent that software is used on computers and computers consume electricity.

Referring to Table 3, residential electricity consumption was relatively more correlated with income than with piracy. Since legitimate consumption is the complement of piracy, this would also imply that residential electricity consumption was relatively more correlated with income than with legitimate consumption.

Table 4 reports tests of endogeneity, using post-change residential electricity consumption per capita as the instrument for post-change income per capita, in two sub-samples. One sub-sample comprised the 15 base countries in which the IDC surveyed consumers and business users to compute norms for software loads and measure computer ownership. The other sub-sample comprised all other countries, for which the IDC somehow projected the norms for software loads or computer ownership. (In Table 4, except as otherwise noted, all variables except year trends were specified in logarithms ([16], pp. 197-200), and all regressions included country fixed effects, and to be conservative, were limited to the balanced panel.)

Referring to Table 4, columns (ii) and (iii), in the first-stage regression of post-change income, the coefficient of post-change electricity consumption was significant, while in the second-stage regression, the coefficient of the residuals from the first-stage regression was significant. This suggests that, over the years 2003-07, following the change in consultant and methodology, income was endogenous. This finding was consistent with the Hypothesis, that the norm for software load or number of computers was projected on the basis of income per capita.

Table 4, column (iv), reports the 2SLS (2-stage least squares) estimates. The coefficient of post-change income was positive and significant (0.098 (±0.029)) while the post-change trend was -0.086 (±0.031). By comparison with the OLS estimate in column (i), the major differences were that the coefficient of post-change income and the post-change trend were about double in magnitude. Apparently, not only did the change in consultant and methodology affect the trend in piracy rates, but it also caused legitimate software consumption to become apparently more sensitive to income per capita. There seems to be no obvious reason for this except the assumptions underlying the BSA consultant's projection of software usage.

The significance of post-change income also helps to confirm that the endogeneity of income was due to the consultant's projection of software usage. An alternative explanation – that both income and usage were related to some omitted factor (e.g., computer literacy or education) – is implausible as there is no obvious reason for the relation between usage and the hypothetical omitted factor to have changed around 2002-03.

By investigating the relation between legitimate consumption and income in the base countries, I can further characterize the impact of the BSA consultant, IDC's projection of software usage. Recall that IDC calculated the software loads from surveys of users in 15 (base) countries, while using projections for the other countries. Accordingly, among the base countries, income should not be endogenous.

Table 4, columns (v) to (viii), reports estimates for the base countries. Referring to column (vii), the first-stage residuals were not significant in the second-stage regression. Consistent with my argument above, post-change income was not endogenous among the base countries. This provides indirect corroborating evidence for the Hypothesis.

Since income is exogenous to legitimate consumption among the base countries, the most appropriate estimate for the base countries would be the OLS regression in column (v). Comparing this estimate for the base countries with the 2SLS regression for the projection countries, reported in column (iii). I make two observations.

- The post-change trend, which was marginally significant at -0.040 (±0.020), among the base countries was more than 2 standard errors smaller than the post-change trend, -0.086 (±0.031), among the projection countries.
- The coefficient of post-change income, which was marginally significant at 0.056 (±0.026) among the base countries, was about 1.5 standard errors smaller than the coefficient, 0.098 (±0.028), among the projection countries.

The disparity between the estimates for the base vis-à-vis projection countries suggests two systematic biases in the IDC's measurement of piracy from 2003 onward. Apparently, the projection of software usage on the basis of income caused (i) a stronger downward trend in legitimate consumption, and (ii) legitimate consumption to be over-sensitive to income.

By (4), the piracy rate is just the complement of the legitimate consumption rate. Supposing that the data on legitimate consumption among the base countries was more reliable, I infer that IDC's projection of software usage resulted in an under-estimate of the trend decrease in the piracy rate and an over-estimate of the sensitivity of piracy to changes in income among the projection countries.

Finally, I note that, among neither the projection countries nor the base countries, did I find any evidence that income was endogenous to published rates of software piracy between 1997-2002. This is consistent with the consultant, IPRC's disclosure that it directly applied U.S. norms for software loads to other countries, without any adjustment ([3], page 11).

5. Conclusions

U.S. government pronouncements and actions as well as many academic studies have taken BSA software piracy statistics at face value. Based on a review of the BSA methodology and empirical analysis, I conclude that a change in the BSA consultant and methodology around 2002-03 had systematic effects on published piracy rates.

- The trend rate of decrease of piracy rates fell from 2.0% points per year to 1.1% points per year, so, raising piracy rates from the levels that would have been implied had they followed the trend before the change.
- Among countries for which the BSA's piracy statistics depended on a projection of software usage, the trend decrease in the piracy rate was under-estimated and the sensitivity of piracy to changes in income was over-estimated.

An important question is the robustness of these findings to the sample of countries, choice of instruments, and inclusion of alternative explanatory variables. In additional tests (unreported for brevity), I have investigated and confirmed the robustness of the results to these issues.

The key direction for future research is to gain access to the BSA methodologies and data so as to better understand the biases in their statistics, and so that future policy and research can be appropriately calibrated. Meanwhile, the central implication of my analysis is that BSA statistics should be used with great caution. Any government pronouncement or action, and academic study using the BSA software piracy statistics should take account of changes in the BSA consultant and methodology. Software piracy statistics for countries for which BSA undertook projections of software usage should be treated with special care.

9. References

[1] Trisha L. Bezmen and Craig A. Depken II, “Influences on software piracy: Evidence from the various United States”, Economics Letters, 90(3), 2006, pp. 356-361.

[2] Business Software Alliance, First Annual BSA and IDC Global Software Piracy Study, 2004.

[3] Business Software Alliance, Eighth Annual BSA Global Software Piracy Study: Trends in Software Piracy, 1994-2002, June 2003.

[4] Ramnath K. Chellappa, Yuanyuan Chen, and Sriram Venkataraman, “An Empirical Examination of Global Software Piracy: Pricing and Policy Implications”, Goizueta Business School, Emory University, Atlanta GA 30322, 2006.

[5] Craig A. Depken, II, and Lee C. Simmons, “Social construct and the propensity for software piracy”, Applied Economics Letters, 11(2), 2004, pp. 97–100.

[6] Justina A.V. Fischer and Antonio Rodríguez Andrés, “Is Software Piracy a Middle Class Crime? Investigating the Inequality-Piracy Channel”, Discussion paper no. 2005-18, Department of Economics, University of St. Gallen, August 2005.

[7] Ram D. Gopal and G. Lawrence Sanders, “International Software Piracy: Analysis of Key Issues and Impacts”, Information Systems Research, 9(4), 1998, pp. 380-397.

[8] Ram D. Gopal and G. Lawrence Sanders, “Global software piracy: you can't get blood out of a turnip”, Communications of the ACM, 43(9), 2000, pp. 82 – 89.

[9] Bryan W. Husted, “The Impact of National Culture on Software Piracy”, Journal of Business Ethics, 26(3), 2000, pp. 197-211.

[10] International Intellectual Property Alliance (IIPA), Press Release: The Copyright Industries in the International Intellectual Property Alliance (IIPA) Submit to USTR their 2007 Report on Piracy in 60 Countries/Territories, Washington, DC, February 17, 2007.

[11] Donald B. Marron and David G. Steel, “Which Countries Protect Intellectual Property? The Case of Software Piracy”, Economic Inquiry, 38(2), 2000, pp. 159-174.

[12] Karsten Olsen, “Counterfeiting and Piracy: Measurement Issues”, Background report for the WIPO/OECD Expert Meeting on Measurement and Statistical Issues, Geneva, 17-18 October 2005.

[13] Christopher J. Robertson, K. Matthew Gilley, Victoria Crittenden, and William F. Crittenden, “An analysis of the predictors of software piracy within Latin America”, Journal of Business Research, 61(6), 2008, pp. 651-656.

[14] Andrés Antonio Rodríguez, “The relationship between copyright software protection and piracy: Evidence from Europe”, European Journal of Law and Economics, 21(1), 2006, pp. 29-52.

[15] Hans van Kranenburg and Annelies Hogenbirk, “Multimedia, Entertainment, and Business Software Copyright Piracy: A Cross-National Study”, Journal of Media Economics, 18(2), 2005, pp. 109-129.

[16] Jeffrey M. Wooldridge, Introductory Econometrics: A Modern Approach, 3d edition, Mason, OH: Thomson Southwestern, 2006.

Table 1. Piracy rates

VARIABLES	(i) Years - all OLS	(ii) Trend - all OLS	(iii) Years - all	(iv) Trend - all	(v) Years - balanced	(vi) Trend - balanced
yr98	-1.756		-1.756***		-1.765***	
	(3.045)		(0.296)		(0.300)	
yr99	-5.098*		-5.098***		-5.160***	
	(3.045)		(0.414)		(0.415)	
yr00	-7.463**		-7.553***		-7.617***	
	(3.035)		(0.594)		(0.599)	
yr01	-8.533***		-8.595***		-8.667***	
	(3.009)		(0.713)		(0.719)	
yr02	-10.579***		-10.641***		-10.716***	
	(3.009)		(0.812)		(0.824)	
yr03	-5.781*		-8.324***		-8.247***	
	(2.968)		(1.035)		(1.032)	
yr04	-4.730		-8.097***		-8.222***	
	(2.925)		(1.141)		(1.152)	
yr05	-4.223		-8.552***		-8.605***	
	(2.898)		(1.177)		(1.193)	
yr06	-5.065*		-9.394***		-9.395***	
	(2.898)		(1.207)		(1.231)	
yr07	-6.758**		-10.568***		-10.519***	
	(2.898)		(1.231)		(1.270)	
Pre-change trend		-2.032***		-2.034***		-2.064***
		(0.484)		(0.155)		(0.160)
Post-change trend		-0.733***		-1.138***		-1.148***
		(0.224)		(0.133)		(0.137)
Constant	65.451***	66.886***	67.302***	68.703***	65.815***	67.279***
	(2.153)	(1.881)	(0.719)	(0.811)	(0.689)	(0.800)
Observations	992	992	992	992	891	891
R-squared	0.019	0.017	0.302	0.285	0.300	0.285
No. of countries			103	103	81	81

*** p<0.01, ** p<0.05, * p<0.1

Robust cluster standard errors in parentheses; regressions (iii)-(vi) included fixed effects.

Table 2. Descriptive statistics

	Units	Source	Observati ons	Mean	Std dev	Min.	Max.
Business software piracy	%	BSA	992	59.98	19.59	20	98
GDP per capita (income)	'000 USD (2000 prices)	World Bank	998	9.331	11.029	0.300	54.178
Residential electricity consumption per capita	kWh per capita	GMID[a]	980	1.280	1.466	0.021	7.955

[a] Euromonitor, Global Market Information Database

Table 3. Correlations

	Piracy	GDP per capita	Residential electricity consumption per capita
Piracy	1.000		
GDP per capita	-0.815	1.000	
Residential electricity consumption per capita	-0.589	0.765	1.000

Table 4. Endogeneity of income

(Dependent variable: Log legitimate consumption rate;
Instrument: Log residential electricity consumption)

	Projection countries				15 base countries			
VARIABLES	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
	OLS	1st stg	2nd stg	2SLS	OLS	1st stg	2nd stg	2SLS
Log GDP per capita	1.309***	-3.237***	1.439***	1.439***	1.357***	-2.847**	1.414**	1.414***
	(0.360)	(0.977)	(0.354)	(0.322)	(0.420)	(1.024)	(0.481)	(0.451)
Log post-change	0.058***		0.098***	0.098***	0.056*		0.074	0.074
GDP per capita	(0.015)		(0.028)	(0.029)	(0.026)		(0.080)	(0.076)
Pre-change trend	0.057***	-0.049**	0.059***	0.059***	0.062**	-0.054***	0.063**	0.063***
	(0.011)	(0.021)	(0.012)	(0.011)	(0.024)	(0.016)	(0.024)	(0.023)
Post-change trend	-0.045**	1.041***	-0.086***	-0.086***	-0.040*	1.029***	-0.058	-0.058
	(0.020)	(0.025)	(0.031)	(0.031)	(0.020)	(0.026)	(0.076)	(0.072)
Log post-change		1.030***				0.771***		
electricity		(0.058)				(0.129)		
1st stage residuals			-0.066**				-0.024	
			(0.027)				(0.080)	
Constant	-7.852**	27.992***	-8.978***		-8.017**	23.732**	-8.493**	
	(3.063)	(8.298)	(3.010)		(3.408)	(8.408)	(3.924)	
Observations	626	626	626	626	139	139	139	139
Number of countries	63	63	63	63	14	14	14	14
R-squared	0.350	0.972	0.366	0.325	0.453	0.972	0.455	0.449

*** p<0.01, ** p<0.05, * p<0.1
Robust cluster standard errors in parentheses;
all regressions included fixed effects

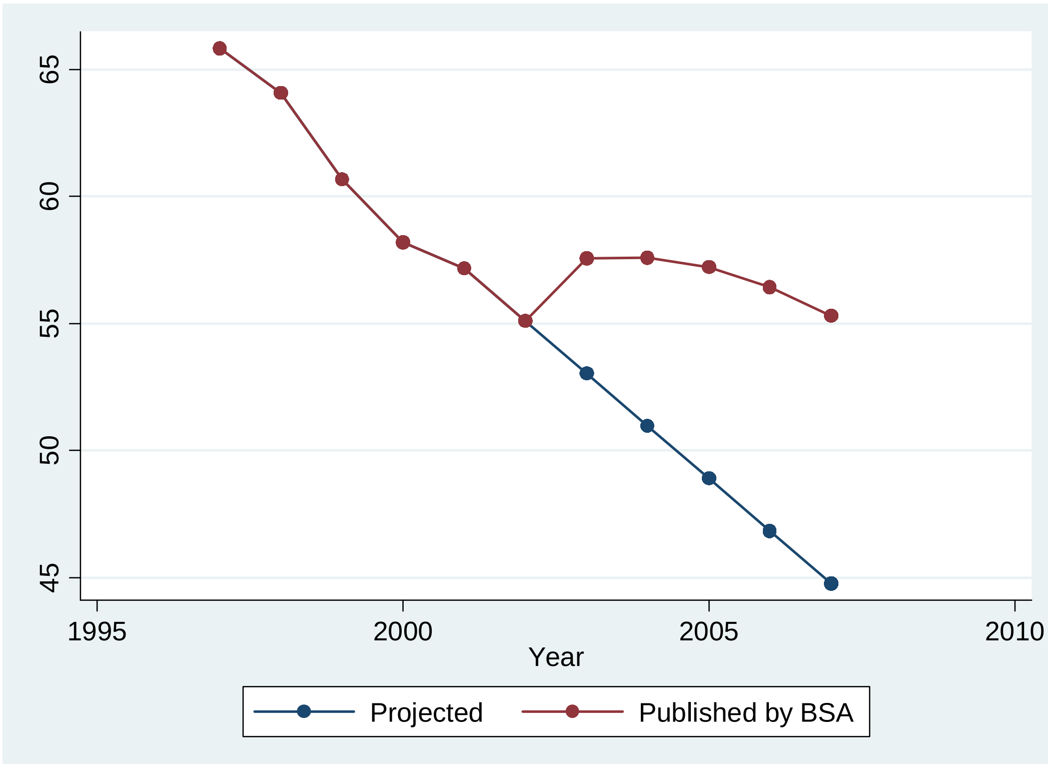

Figure 1. Average piracy rate

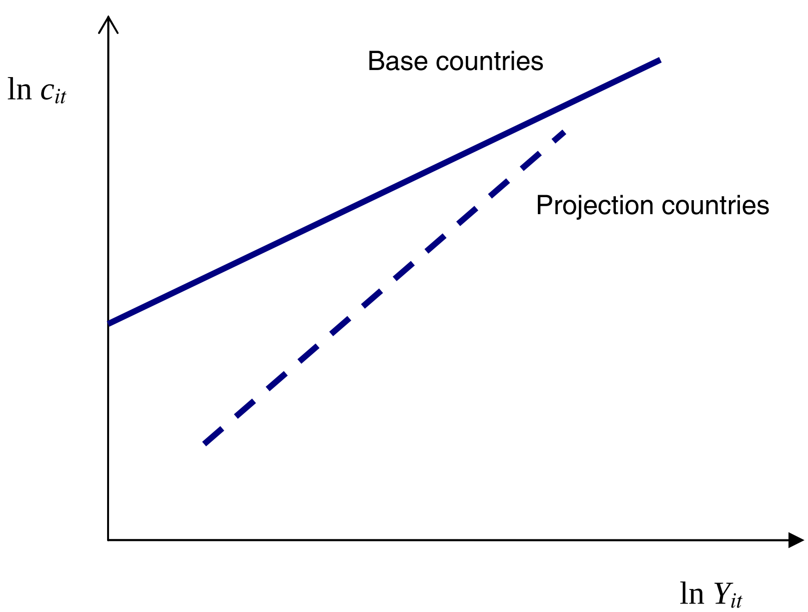

Figure 2. Legitimate consumption rate and income